Multi-Modal Domain Adaptation for Fine-Grained Action Recognition

Jonathan Munro
University of Bristol
jonathan.munro@bristol.ac.uk

Dima Damen
University of Bristol
dima.damen@bristol.ac.uk

Abstract

Fine-grained action recognition datasets exhibit environmental bias, where multiple video sequences are captured from a limited number of environments. Training a model in one environment and deploying in another results in a drop in performance due to an unavoidable domain shift. Unsupervised Domain Adaptation (UDA) approaches have frequently utilised adversarial training between the source and target domains. However, these approaches have not explored the multi-modal nature of video within each domain. In this work we exploit the correspondence of modalities as a self-supervised alignment approach for UDA in addition to adversarial alignment (Fig. 1).

We test our approach on three kitchens from our large-scale dataset, EPIC-Kitchens [8], using two modalities commonly employed for action recognition: RGB and Optical Flow. We show that multi-modal self-supervision alone improves the performance over source-only training by 2.4% on average. We then combine adversarial training with multi-modal self-supervision, showing that our approach outperforms other UDA methods by 3%.

1. Introduction

Fine-grained action recognition is the problem of recognising actions and interactions such as "cutting a tomato" or "tightening a bolt" compared to coarse-grained actions such as "preparing a meal". This has a wide range of applications in assistive technologies in homes as well as in industry. Supervised approaches rely on collecting a large number of labelled examples to train discriminative models. However, due to the difficulty in collecting and annotating such fine-grained actions, many datasets collect long untrimmed sequences. These contain several fine-grained actions from a single [43, 50] or few [8, 47] environments.

Figure 2 shows the recent surge in large-scale fine-grained action datasets. Two approaches have been attempted to achieve scalability: crowd-sourcing scripted actions [17, 46, 47], and long-term collections of natural interactions in homes [8, 37, 43]. While the latter offers more realistic videos, many actions are collected in only a few environments. This leads to learned representations which do not generalise well [53].

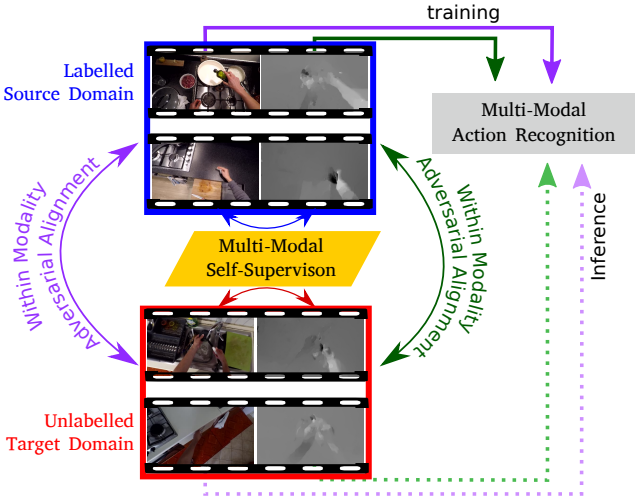


Figure 1: Our proposed UDA approach for multi-modal action recognition. Improved target domain performance is achieved via multi-modal self-supervision on source and target domains simultaneously, jointly optimised with multiple domain discriminators, one per-modality.

Transferring a model learned on a labelled source domain to an unlabelled target domain is known as Unsupervised Domain Adaptation (UDA). Recently, significant attention has been given to deep UDA in other vision tasks [14, 15, 32, 33, 51, 55]. However, very few works have attempted deep UDA for video data [7, 19]. Surprisingly, none have tested on videos of fine-grained actions and all these approaches only consider video as images (*i.e.* RGB modality). This is in contrast with self-supervised approaches that have successfully utilised multiple modalities within video when labels are not present during training [1].

Up to our knowledge, no prior work has explored the **multi-modal** nature of video data for UDA in action recognition. We summarise our contributions as follows:

- We show that multi-modal self-supervision, applied to both source and unlabelled target data, can be used for

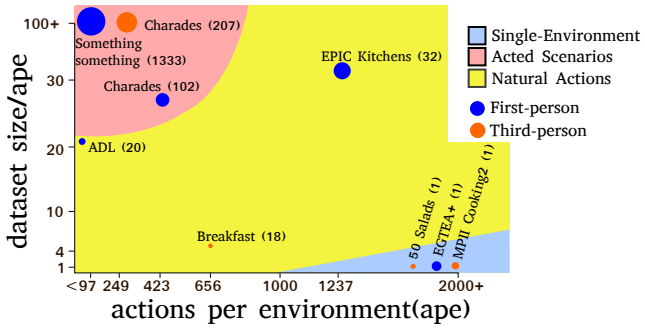


Figure 2: Fine-grained action datasets [8, 17, 26, 28, 38, 42, 46, 47, 50], *x-axis:* number of action segments per environment (ape), *y-axis:* dataset size divided by ape. EPIC-Kitchens [8] offers the largest ape relative to its size.

domain adaptation in video.

- We propose a multi-modal UDA strategy, which we name MM-SADA, to adapt fine-grained action recognition models to unlabelled target environments, using both adversarial alignment and multi-modal self-supervision.
- We test our approach on three domains from EPIC-Kitchens [8], trained end-to-end using I3D [6], and provide the first benchmark of UDA for fine-grained action recognition. Our results show that MM-SADA outperforms source-only generalisation as well as alternative domain adaptation strategies such as batch-based normalisation [29], distribution discrepancy minimisation [32] and classifier discrepancy [45].

2. Related Works

This section discusses related literature starting with general UDA approaches, then supervised and self-supervised learning for action recognition, concluding with works on domain adaptation for action recognition.

Unsupervised Domain Adaptation (UDA) outside of Action Recognition. UDA has been extensively studied for vision tasks including object recognition [14, 15, 32, 33, 51, 55], semantic segmentation [18, 60, 65] and person re-identification [10, 49, 62]. Typical approaches adapt neural networks by minimising a discrepancy measure [15, 51], thus matching mid-level representations of source and target domains. For example, Maximum Mean Discrepancy (MMD) [15, 32, 33] minimises the distance between the means of the projected domain distributions in Reproducing Kernel Hilbert Space. More recently, domain adaptation has been influenced by adversarial training [14, 55]. Simultaneously learning a domain discriminator, whilst maximising its loss with respect to the feature extractor, minimises the domain discrepancy between source and target. In [55], a GAN-like loss function allows separate weights for source and target domains, while in [14] shared weights are used, efficiently removing domain specific information by inverting the gradient produced by the domain discriminator with a Gradient Reversal Layer (GRL).

Utilising multiple modalities (image and audio) for UDA has been recently investigated for bird image retrieval [39]. Multiple adversarial discriminators are trained on a single modality as well as mid-level fusion and a cross-modality attention is learnt. The work shows the advantages of multi-modal domain adaptation in contrast to single-modality adaptation, though in their work both modalities demonstrate similar robustness to the domain shift.

Very recently, self-supervised learning has been proposed as a domain adaptation approach [5, 52]. In [5], it is used as an auxiliary task, by jigsaw-shuffling image patches and predicting their permutations over multiple source domains. In [52], self-supervision was shown to replace adversarial training using tasks such as predicting rotation and translation for object recognition. In the same work, self-supervision was shown to benefit adversarial training when jointly trained for semantic segmentation. Both works only use a single image. Our work utilises the multiple modalities offered by video, showing that self-supervision can be used to adapt action recognition models to target domains.

Supervised Action Recognition. Convolutional networks are state of the art for action recognition, with the first seminal works using either 3D [20] or 2D convolutions [22]. Both utilise a single modality—appearance information from RGB frames. Simonyan and Zisserman [48] address the lack of motion features captured by these architectures, proposing two-stream late fusion that learns separate features from the Optical Flow and RGB modalities, outperforming single modality approaches.

Following architectures have focused on modelling longer temporal structure, through consensus of predictions over time [30, 58, 63] as well as inflating CNNs to 3D convolutions [6], all using the two-stream approach of late-fusing RGB and Flow. The latest architectures have focused on reducing the high computational cost of 3D convolutions [12, 21, 61], yet still show improvements when reporting results of two-stream fusion [61].

Self-supervision for Action Recognition. Self-supervision methods learn representations from the temporal [13, 59] and multi-modal structure of video [1, 25], leveraging pre-training on a large corpus of unlabelled videos. Methods exploiting the temporal consistency of video have predicted the order of a sequence of frames [13] or the arrow of time [59]. Alternatively, the correspondence between multiple modalities has been exploited for self-supervision, particularly with audio and RGB [1, 25, 35]. Works predicted if modalities correspond or are synchronised. We test both approaches for self-supervision in our UDA approach.

Domain Adaptation for Action Recognition. Of the several domain shifts in action recognition, only one has received significant research attention, that is the problem

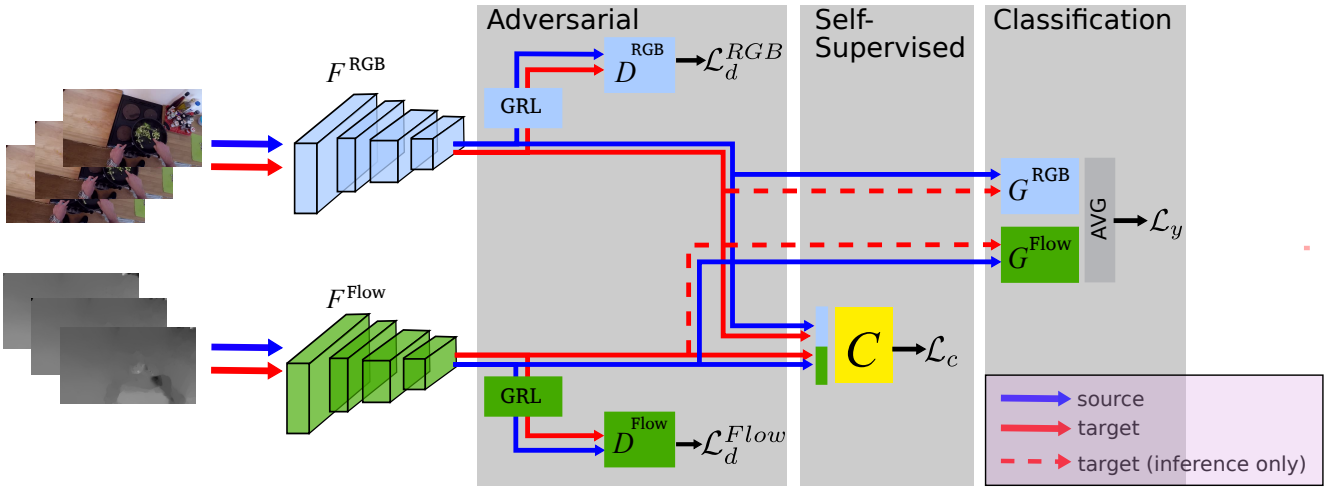


Figure 3: Proposed architecture: feature extractors F^{RGB} and F^{Flow} are shared for both target and source domains. Domain Discriminators, D^{RGB} and D^{Flow}, are applied to each modality. Self-supervised correspondence of modalities, C, is trained from both source and unlabelled target data. Classifiers, G^{RGB} and G^{Flow} are trained using source domain examples only from the average pooled classification scores of each modality. During inference, multimodal target data is classified.

of cross-viewpoint (or viewpoint-invariant) action recognition [24, 27, 31, 40, 46]. These works focus on adapting to the geometric transformations of a camera but do little to combat other shifts, like changes in environment. Works utilise supervisory signals such as skeleton or pose [31] and corresponding frames from multiple viewpoints [24, 46]. Recent works have used GRLs to create a view-invariant representation [27]. Though several modalities (RGB, flow and depth) have been investigated, these were aligned and evaluated independently.

On the contrary, UDA for changes in environment has received limited recent attention. Before deep-learning, UDA for action recognition used shallow models to align source and target distributions of handcrafted features [4, 11, 64]. Three recent works attempted deep UDA [7, 19, 36]. These apply GRL adversarial training to C3D [54], TRN [63] or both [36] architectures. Jamal *et al.*'s approach [19] outperforms shallow methods that use subspace alignment. Chen *et al.* [7] show that attending to the temporal dynamics of videos can improve alignment. Pan *et al.* [36] use a cross-domain attention module, to avoid uninformative frames. Two of these works use RGB only [7, 19] while [36] reports results on RGB and Flow, however, modalities are aligned independently and only fused during inference. The approaches [7, 19, 36] are evaluated on 5-7 pairs of domains from subsets of coarse-grained action recognition and gesture datasets, for example aligning UCF [41] to Olympics [34]. We evaluate on 6 pairs of domains. Compared to [19], we use 3.8× more training and 2× more testing videos.

The EPIC-Kitchens [8] dataset for fine-grained action recognition released two distinct test sets—one with seen and another with unseen/novel kitchens. In the 2019 challenges report, all participating entries exhibit a drop in action recognition accuracy of 12-20% when testing their models on novel environments compared to seen environments [9]. Up to our knowledge, no previous effort applied UDA on this or any fine-grained action dataset.

In this work, we present the first approach to multi-modal UDA for action recognition, tested on fine-grained actions. We combine adversarial training on multiple modalities with a modality correspondence self-supervision task. This utilises the differing robustness to domain shifts between the modalities. Our method is detailed next.

3. Proposed Method

This section outlines our proposed action recognition domain adaptation approach, which we call *Multi-Modal Self-Supervised Adversarial Domain Adaptation (MM-SADA)*. In Fig. 3, we present an overview of MM-SADA, visualised for action recognition using two modalities: RGB and Optical Flow. We incorporate a self-supervision alignment classifier, C, that determines whether modalities are sampled from the same or different actions to learn modality correspondence. This takes in the concatenated features from both modalities, without any labels. Learning the correspondence on source and target encourages features that generalise to both domains. Aligning the domain statistics is achieved by adversarial training, with a domain discriminator per modality that predicts the domain. A Gradient Reversal layer (GRL) reverses and backpropagates the gradient to the features. Both alignment techniques are trained on source and unlabelled target data whereas the action classifier is only trained with labelled source data.

We next detail MM-SADA, generalised to any two or

more modalities. We start by revisiting the problem of *domain adaptation* and outlining multi-stream late fusion, then we describe our adaptation approach.

3.1. Unsupervised Domain Adaptation (UDA)

A domain is a distribution over the input population $\mathbf{X}$ and the corresponding label space $\mathbf{Y}$. The aim of supervised learning, given labelled samples $\{(x, y)\}$, is to find a representation, $G(\cdot)$, over some learnt features, $F(\cdot)$, that minimises the empirical risk, $E_{\mathbf{S}}[\mathcal{L}_y(G(F(x)), y)]$. The empirical risk is optimised over the labelled source domain, $\mathbf{S} = \{X^s, Y^s, \mathcal{D}^s\}$, where $\mathcal{D}^s$ is a distribution of source domain samples. The goal of domain adaptation is to minimise the risk on a target domain, $\mathbf{T} = \{X^t, Y^t, \mathcal{D}^t\}$, where the distributions in the source and target domains are distinct, $\mathcal{D}^s \neq \mathcal{D}^t$. In UDA, the label space Y^t is unknown, thus methods minimise both the source risk and the distribution discrepancy between the source and target domains [3].

3.2. Multi-modal Action Recognition

When the input is multi-modal, *i.e.* $X = (X^1, \cdots, X^M)$ where X^m is the m^{th} modality of the input, fusion of modalities can be employed. Most commonly, late fusion is implemented, where we sum prediction scores from modalities and backpropagate the error to all modalities, *i.e.*:

$$\mathcal{L}_y = \sum_{x \in \{\mathbf{S}\}} -y \log P(x)$$
$$\text{where: } P(x) = \sigma\Big(\sum_{m=1}^{M} G^m(F^m(x^m))\Big) \quad (1)$$

where G^m is the modality's task classifier, and F^m is the modality's learnt feature extractor. The consensus of modality classifiers is trained by a cross entropy loss, $\mathcal{L}_y$, between the task label, y, and the prediction, $P(x)$. σ is defined as the softmax function. Training for classification expects the presence of labels and thus can only be applied to the labelled source input.

3.3. Within-Modal Adversarial Alignment

Both generative and discriminative adversarial approaches have been proposed for bridging the distribution discrepancy between source and target domains. Discriminative approaches are most appropriate with high-dimensional input data present in video. Generative adversarial requires a huge amount of training data and temporal dynamics are often difficult to reconstruct. Discriminative methods train a discriminator, $D(\cdot)$, to predict the domain of an input (*i.e.* source or target), from the learnt features, $F(\cdot)$. By maximising the discriminator loss, the network learns a feature representation that is invariant to both domains.

For aligning multi-modal video data, we propose using a domain discriminator per modality that penalises domain specific features from each modality's stream. Aligning modalities separately avoids the easier solution of the network focusing only on the less robust modality in classifying the domain. Each separate domain discriminator, D^m, is thus used to train the modality's feature representation F^m. Given a binary domain label, d, indicating if an example $x \in \mathbf{S}$ or $x \in \mathbf{T}$, the domain discriminator, for modality m, is defined as,

$$\mathcal{L}_d^m = \sum_{x \in \{\mathbf{S},\mathbf{T}\}} -d \log(D^m(F^m(x))) - (1-d) \log(1 - D^m(F^m(x))) \quad (2)$$

3.4. Multi-Modal Self-Supervised Alignment

Prior approaches to domain adaptation have mostly focused on images and thus have not explored the multi-modal nature of the input data. Videos are multi-modal, where corresponding modalities are present in both source and target. We thus propose a multi-modal self-supervised task to align domains. Multi-modal self-supervision has been successfully exploited as a pretraining strategy [1, 2]. However, we show that self-supervision for both source and target domains can also align domains.

We learn the temporal correspondence between modalities as a self-supervised binary classification task. For positive examples, indicating that modalities correspond, we sample modalities from the same action. These could be from the same time, or different times within the same action. For negative examples, each modality is sampled from a different action. The network is thus trained to determine if the modalities correspond. This is optimised over both domains. A self-supervised correspondence classifier head, C, is used to predict if modalities correspond. This shares the same modality feature extractors, F^m, as the action classifier. It is important that C is as shallow as possible so that most of the self-supervised representation is learned in the feature extractors. Given a binary label defining if modalities correspond, c, for each input, x, and concatenated features of the multiple modalities, we calculate the multi-modal self-supervision loss as follows:

$$\mathcal{L}_c = \sum_{x \in \{\mathbf{S},\mathbf{T}\}} -c \log C(F^0(x), ..., F^M(x)) \quad (3)$$

3.5. Proposed MM-SADA

We define the Mutli-Modal Self-Supervised Adversarial Domain Adaptation (MM-SADA) approach as follows. The classification loss, $\mathcal{L}_y$, is jointly optimised with the adversarial and self-supervised alignment losses. The within-modal adversarial alignment is weighted by λ_d, and the multi-modal self-supervised alignment is weighted by λ_c. Optimising both alignment strategies achieves benefits in

matching source and target statistics and learning cross-modal relationships transferable to the target domain.

$$\mathcal{L} = \mathcal{L}_y + \lambda_d \sum_m \mathcal{L}_d^m + \lambda_c \mathcal{L}_c \quad (4)$$

Note that the first loss $\mathcal{L}_y$ is only optimised for labelled source data, while the alignment losses $\forall m : \mathcal{L}_d^m$ and $\mathcal{L}_c$ are optimised for both unlabelled source and target data.

4. Experiments and Results

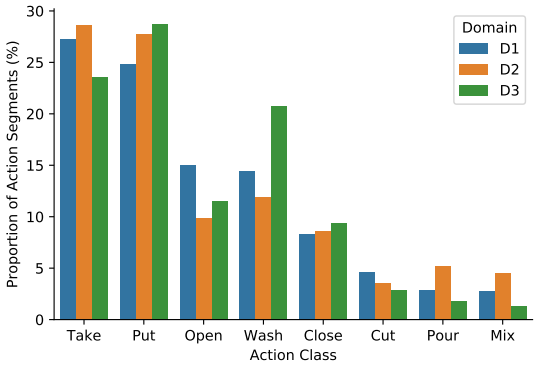
This section first discusses the dataset, architecture, and implementation details in Sec. 4.1. We compare against baseline methods noted in Sec. 4.2. Results are presented in Sec. 4.3, followed by an ablation study of the method's components in Sec. 4.4 and qualitative results including feature space visualisations in Sec. 4.5.

4.1. Implementation Details

Dataset. Our previous work, EPIC Kitchens [8], offers a unique opportunity to test domain adaptation for fine-grained action recognition, as it is recorded in 32 environments. Similar to previous works for action recognition [14, 19], we evaluate on pairs of domains. We select the three largest kitchens, in number of training action instances, to form our domains. These are P01, P22, P08, which we refer to as D1, D2 and D3, respectively (Fig. 4).

We analyse the performance for the 8 largest action classes: ('put', 'take', 'open', 'close', 'wash', 'cut', 'mix', and 'pour'), which form 80% of the training action segments for these domains. This ensures sufficient examples per domain and class, without balancing the training set. The label imbalance of these 8 classes is depicted in Fig. 4 (middle) which also shows the differing distribution of classes between the domains. Most domain adaptation works evaluate on balanced datasets [14, 16, 44] with few using imbalanced datasets [57]. EPIC-Kitchens has a large class imbalance offering additional challenges for domain adaptation. The number of action segments in each domain are specified in Fig. 4 (bottom), where a segment is a labeled start/end time, with an action label.

Architecture. We train all our models end-to-end. We use the inflated 3D convolutional architecture (I3D) [6] as our backbone for feature extraction, one per modality (F^m). In this work, F convolves over a temporal window of 16 frames. In training, a single temporal window is randomly sampled from within the action segment each iteration. In testing, as in [58], we use an average over 5 temporal windows, equidistant within the segment. We use the RGB and Optical Flow frames provided publicly [8]. The output of F is the result of the final average pooling layer of I3D, with 1024 dimensions. G is a single fully connected layer with softmax activation to predict class labels. Each domain discriminator D^m is composed of 2 fully connected



Domain	D1	D2	D3
Ref. EPIC Kitchen	P08	P01	P22
Training Action Segments	1543	2495	3897
Test Action Segments	435	750	974

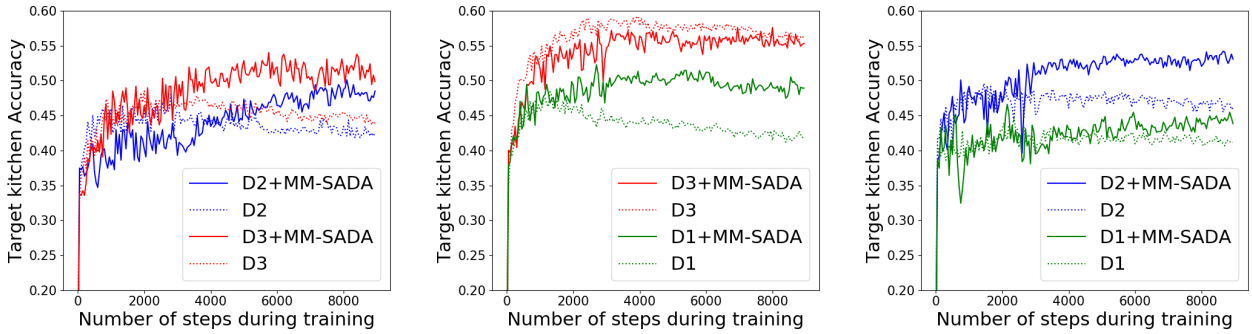
Figure 4: **Top:** Three kitchens from EPIC-Kitchens selected as domains to evaluate our method **Middle:** Class distribution per domain, for the 8 classes in legend. **Bottom:** Number of action segments per domain.

layers with a hidden layer of 100 dimensions and a ReLU activation function. A dropout rate of 0.5 was used on the output of F and $1e-7$ weight decay for all parameters. Batch normalisation layers are used in F^m and are updated with target statistics for testing, as in AdaBN [29]. We apply random crops, scale jitters and horizontal flips for data augmentation as in [58]. During testing only center crops are used. The self-supervised correspondence function C (Eq. 3) is implemented as 2 fully connected layers of 100 dimensions and a ReLU activation function. The features from both modalities are concatenated along the channel dimension as input to C.

Training and Hyper-parameter Choice. We train using the Adam optimiser [23] in two stages. First the network is trained with only the classification and self supervision losses $\mathcal{L}_y + \lambda_c \mathcal{L}_c$ at a learning rate of $1e-2$ for 3K iterations. Then, the overall loss function (Eq. 4) is optimised, applying the domain adversarial losses $\mathcal{L}_d^m$, and reducing the learning rate to $2e-4$ for a further 6K steps.

	D2→ D1	D3→ D1	D1→ D2	D3→ D2	D1→ D3	D2→ D3	Mean
MM Source-only	42.5	44.3	42.0	**56.3**	41.2	46.5	45.5
AdaBN [29]	44.6	47.8	47.0	54.7	40.3	48.8	47.2
MMD [32]	43.1	48.3	46.6	55.2	39.2	48.5	46.8
MCD [45]	42.1	47.9	46.5	52.7	43.5	51.0	47.3
MM-SADA	**48.2** ▲+5.7	**50.9**▲+6.6	**49.5**▲+7.5	56.1▼-0.2	**44.1**▲+2.9	**52.7** ▲+6.3	**50.3** ▲+4.8
Supervised target	62.8	62.8	71.7	71.7	74.0	74.0	69.5

Table 1: Top-1 Accuracy on the target domain, for our proposed MM-SADA, compared to different alignment approaches. On average, we outperform the source-only performance by 4.8%.



(a) Target D1 (b) Target D2 (c) Target D3

Figure 5: Accuracy on target during training epochs. Solid line is MM-SADA and dotted line is source-only performance.

The self-supervision hyper-parameter, $\lambda_c = 5$ was chosen by observing the performance on the labelled **source domain** only, *i.e.* this has not been optimised for the target domain. Note that while training with self-supervision, half the batch contains corresponding modalities and the other non-corresponding modalities. Only source examples with corresponding modalities are used to train for action classification. The domain adversarial hyper-parameter, $\lambda_d = 1$, was chosen arbitrarily; we show that the results are robust to some variations in this hyper-parameter in an ablation study. Batch size was set to 128, split equally for source and target samples. On average, training takes 9 hours on an NVIDIA DGX-1 with 8 V100 GPUs.

4.2. Baselines

For all results, we report the top-1 target accuracy averaged over the last 9 epochs of training, for robustness. We first evaluate the impact of domain shift between source and target by testing using a multi-modal source-only model (MM source-only), trained with no access to unlabelled target data. Additionally, we compare to 3 baselines for unsupervised domain adaptation as follows:

– *AdaBN* [29]: Batch Normalisation layers are updated with target domain statistics.

– *Maximum Mean Discrepancy (MMD)*: The multiple kernel implementation of the commonly used domain discrepancy measure MMD is used as a baseline [32]. This directly replaces the adversarial alignment with separate discrepancy measures applied to individual modalities.

– *Maximum Classifier Discrepancy (MCD)* [45]: Alignment through classifier disagreement is used. We use two multi-modal classification heads as separate classifiers. The classifiers are trained to maximise prediction disagreement on the target domain, implemented as L1 loss, finding examples out of support from the source domain. We use a GRL to optimise the feature extractors.

Additionally, as an upper limit, we also report the supervised target domain results. This is a model trained on labelled target data and only offers an understanding of the upper limit for these domains. We highlight these results in the table to avoid confusion.

4.3. Results

First we compare our proposed method MM-SADA to the various domain alignment techniques in Table 1. We show that our method outperforms batch-based [29] (by 3.1%), classifier discrepancy [45] (by 3%) and discrepancy minimisation alignment [32] (by 3.5%) methods. The improvement is consistent for all pairs of domains. Additionally, it significantly improves on the source-only baseline by up to 7.5% in 5 out of 6 cases. For a single case, $D3 \rightarrow D2$, all baselines under-perform compared to

	λ_d	λ_c	D2→ D1	D3→ D1	D1→ D2	D3→ D2	D1→ D3	D2→ D3	Mean
Source-only	0	0	42.5	44.3	42.0	56.3	41.2	46.5	45.5
MM-SADA (Self-Supervised only)	0	5	41.8	49.7	47.7	**57.4**	40.3	50.6	47.9▲+2.4
MM-SADA (Adversarial only)	1	0	46.5	51.0	50.0	53.7	43.5	51.5	49.4▲+3.9
MM-SADA (Adversarial only)	0.5	0	46.9	50.2	50.2	53.6	**44.7**	50.8	49.4▲+3.9
MM-SADA	0.5	5	45.8	**52.1**	**50.4**	56.9	43.5	51.9	50.1▲+4.6
MM-SADA	1	5	**48.2**	50.9	49.5	56.1	44.2	**52.7**	**50.3**▲+4.8

Table 2: Ablation of our method, showing the contribution of the various loss functions (Eq 4). When $\lambda_d = 0$, modality adversarial is not utilised. When $\lambda_c = 0$, self-supervision is not utilised.

	D2→ D1	D3→ D1	D1→ D2	D3→ D2	D1→ D3	D2→ D3	Mean
RGB source-only	37.0	36.3	36.1	44.8	36.6	33.6	37.4
RGB (Adversarial-only)	37.8	41.1	45.7	45.1	38.1	41.2	41.5
RGB (MM-SADA)	41.7	42.1	45.0	48.4	39.7	46.1	43.9
Flow source-only	44.6	44.4	52.2	54.0	41.1	50.0	47.7
Flow (Adversarial-only)	45.5	46.8	51.1	54.6	44.2	47.1	48.2
Flow (MM-SADA)	45.0	45.7	49.0	58.9	44.8	52.1	49.3

Table 3: Ablation of our method on individual modalities, reporting predictions from each modality stream, before late fusion. Note that we still use both modalities for self-supervision. MM-SADA provides improvements for both modalities.

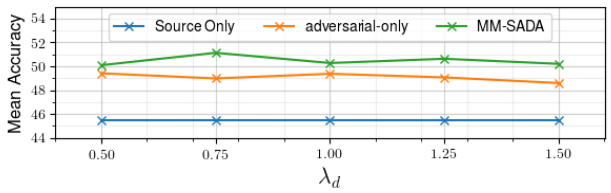


Figure 6: Robustness of the average top-1 accuracy over all pairs of domains for various λ_d on the target domain.

source-only. Ours has a slight drop (-0.2%) but outperforms other alignment approaches. We will revisit this case in the ablation study.

Figure 5 shows the top-1 accuracy on the target during training (solid lines) vs source-only training without domain adaptation (dotted lines). Training without adaptation has consistently lower accuracy, except for our failure case $D3 \rightarrow D2$, showing the stability and robustness of our method during training, with minimal fluctuations due to stochastic optimisation on batches. This is essential for UDA as no target labels can be used for early stopping.

4.4. Ablation Study

Next, we compare the individual contributions of different components of MM-SADA. We report these results in Table 2. The self-supervised component on its own gives a 2.4% improvement over no adaption. This shows that self-supervision can learn features common to both source and target domains, adapting the domains. Importantly, this on average outperforms the three baselines in Table 1. Adversarial alignment per modality gives a further 2.4% improvement as this encourages the source and target distributions to overlap, removing domain specific features from each modality. Compared to adversarial alignment only, our method improves in 5 of the 6 domains and by up to 3.2%.

For the single pair noted earlier, $D3 \rightarrow D2$, self-supervision alone outperforms source-only and all other methods reported in Table 1 by 1.1%. However when combined with domain adaptation using $\lambda_d = 1$, the overall performance of MM-SADA reported in Table 1 cannot beat the baseline. In Table 2, we show that when halving the contribution of adversarial component to $\lambda_d = 0.5$, MM-SADA can achieve 56.9% outperforming the source-only baseline. Therefore self-supervision can improve performance where marginal alignment domain adaptation techniques fail.

Figure 6 plots the performance of MM-SADA as λ_d changes. Note that λ_c can be chosen by observing the performance of self-supervision on source-domain labels, while λ_d requires access to target data. We show that our approach is robust to various values of λ_d, with even higher accuracy at $\lambda_d = 0.75$ than those reported in Table 2.

Table 3 shows the impact of our method on the performance of the modalities individually. Predictions are taken from each modality separately before late fusion. RGB, the less robust modality, benefits most from MM-SADA, improving over source-only by 6.5% on average, whereas Flow improves by 1.6%. The inclusion of multi-modal self-supervision provides 2.4% and 1.1% improvements for

Self-Supervision	D2→ D1	D3→ D1	D1→ D2	D3→ D2	D1→ D3	D2→ D3	**Mean**
Sync.	44.2	50.2	48.0	54.6	41.0	49.4	**47.9**
Seg. Corr.	41.8	49.7	47.7	57.4	40.3	50.6	**47.9**

Table 4: Comparision of two self-supervision tasks for modality correspondence: determining modality synchrony vs. determining whether modality samples come from the same segment. The two approaches perform comparably on average.

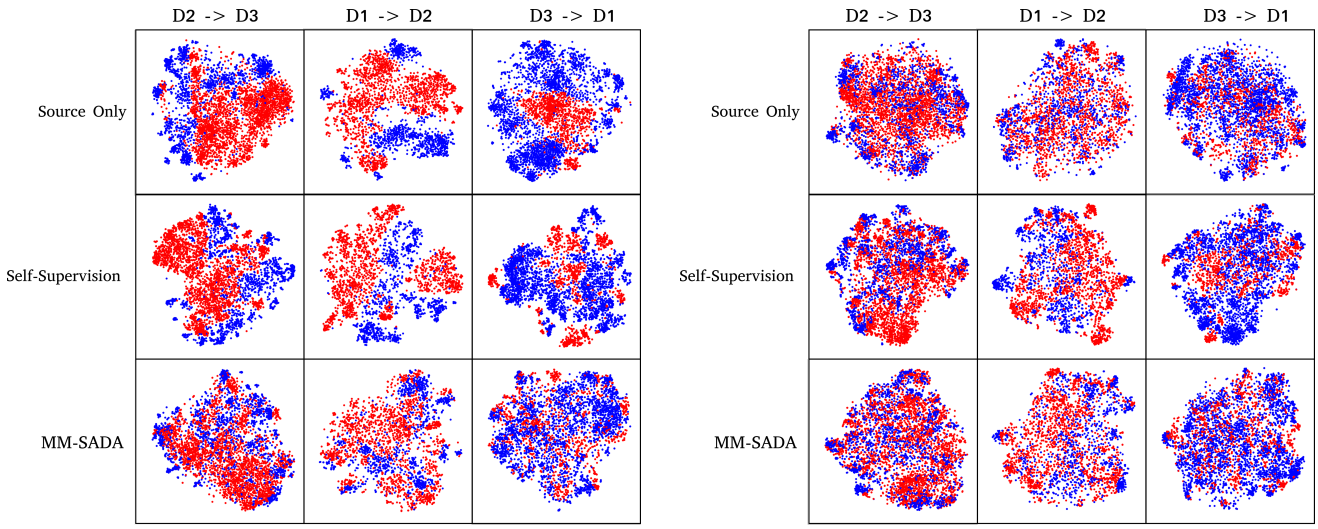


Figure 7: t-SNE plots of RGB (left) and Flow (right) feature spaces produced by source-only, self-supervised alignment and our proposed model MM-SADA. Target is shown in red and source in blue. Our method better aligns both modalities.

RGB and Flow, compared to only using adversarial alignment. This shows the benefit of employing self-supervision from multiple modalities during alignment.

We also compare two approaches for multi-modal self-supervision in Table 4. The first, which has been used to report all results above, learns the correspondence of RGB and Flow within the same action segment. We refer to this as *'Seg. Corr.'*. The second learns the correspondence only from time-synchronised RGB and Flow data, which we call *'Sync'*. The two approaches are comparable in performance overall, with no difference on average over the domain pairs. This shows the potential to use a number of multi-modal self-supervision tasks for alignment.

4.5. Qualitative Results

Figure 7 shows the t-SNE [56] visualisation of the RGB (left) and Flow (right) feature spaces F^m. Several observations are worth noting from this figure. First, Flow shows higher overlap between source and target features pre-alignment (first row). This shows that Flow is more robust to environmental changes. Second, self-supervision alone (second row) changes the feature space by separating the features into clusters, that are potentially class-relevant. This is most evident for $D3 \rightarrow D1$ on the RGB modality (second row third column). However, alone this feature space still shows domain gaps, particularly for RGB features. Third, our proposed MM-SADA (third row) aligns the marginal distributions of source and target domains.

5. Conclusion and Future Work

We proposed a multi-modal domain adaptation approach for fine-grained action recognition utilising multi-modal self-supervision and adversarial training per modality. We show that the self-supervision task of predicting the correspondence of multiple modalities is an effective domain adaptation method. On its own, this can outperform domain alignment methods [32, 45], by jointly optimising for the self-supervised task over both domains. Together with adversarial training, the proposed approach outperforms non-adapated models by 4.8%. We conclude that aligning individual modalities whilst learning a self-supervision task on source and target domains can improve the ability of action recognition models to transfer to unlabelled environments.

Future work will focus on utilising more modalities, such as audio, to aid domain adaptation as well as exploring additional self-supervised tasks for adaptation, trained individually as well as for multi-task self-supervision.

Acknowledgement Research supported by EPSRC LOCATE (EP/N033779/1) and EPSRC Doctoral Training Partnershipts (DTP). The authors acknowledge and value the use of the ESPRC funded Tier 2 facility, JADE.

References

[1] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *International Conference on Computer Vision (ICCV)*, 2017. 1, 2, 4

[2] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *European Conference on Computer Vision (ECCV)*, 2018. 4

[3] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems (NIPS)*, 2006. 4

[4] Liangliang Cao, Zicheng Liu, and Thomas S Huang. Cross-dataset action detection. In *Computer Vision and Pattern Recognition (CVPR)*, 2010. 3

[5] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[6] Joao Carreira and Andrew Zisserman. *Quo Vadis*, action recognition? A new model and the Kinetics dataset. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 5

[7] Min-Hung Chen, Zsolt Kira, Ghassan AlRegib, Jaekwon Yoo, Ruxin Chen, and Jian Zheng. Temporal attentive alignment for large-scale video domain adaptation. In *International Conference on Computer Vision (ICCV)*, October 2019. 1, 3

[8] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, and Will Price. Scaling egocentric vision: The EPIC-Kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018. 1, 2, 3, 5

[9] Dima Damen, Will Price, Evangelos Kazakos, Giovanni Maria Farinella, and Antonino Furnari. EPIC-KITCHENS - 2019 challenges report. *Online Report*, 2019. 3

[10] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[11] N Faraji Davar, Teofilo de Campos, David Windridge, Josef Kittler, and William Christmas. Domain adaptation in the context of sport video action recognition. In *Domain Adaptation Workshop, in conjunction with NIPS*, 2011. 3

[12] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. SlowFast networks for video recognition. In *International Conference on Computer Vision (ICCV)*, October 2019. 2

[13] Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould. Self-supervised video representation learning with odd-one-out networks. In *International Conference on Computer Vision (ICCV)*, 2017. 2

[14] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016. 1, 2, 5

[15] Muhammad Ghifary, W Bastiaan Kleijn, and Mengjie Zhang. Domain adaptive neural networks for object recognition. In *Pacific Rim International Conference on Artificial Intelligence*, 2014. 1, 2

[16] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *Computer Vision and Pattern Recognition (CVPR)*, 2012. 5

[17] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, and Moritz Mueller-Freitag. The "Something Something" video database for learning and evaluating visual common sense. In *International Conference on Computer Vision (ICCV)*, 2017. 1, 2

[18] Haoshuo Huang, Qixing Huang, and Philipp Krahenbuhl. Domain transfer through deep activation matching. In *European Conference on Computer Vision (ECCV)*, September 2018. 2

[19] Arshad Jamal, Vinay P Namboodiri, Dipti Deodhare, and KS Venkatesh. Deep domain adaptation in action space. In *British Machine Vision Conference (BMVC)*, 2018. 1, 3, 5

[20] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3D convolutional neural networks for human action recognition. *Pattern Analysis and Machine Intelligence (PAMI)*, 35(1):221–231, 2013. 2

[21] Boyuan Jiang, MengMeng Wang, Weihao Gan, Wei Wu, and Junjie Yan. STM: Spatiotemporal and motion encoding for action recognition. In *International Conference on Computer Vision (ICCV)*, October 2019. 2

[22] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR)*, 2014. 2

[23] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 5

[24] Yu Kong, Zhengming Ding, Jun Li, and Yun Fu. Deeply learned view-invariant features for cross-view action recognition. *Transactions on Image Processing*, 26(6), 2017. 3

[25] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *Advances in Neural Information Processing Systems (Neurips)*, 2018. 2

[26] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Computer Vision and Pattern Recognition (CVPR)*, 2014. 2

[27] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan Kankanhalli. Unsupervised learning of view-invariant action representations. In *Advances in Neural Information Processing Systems (Neurips)*, 2018. 3

[28] Yin Li, Miao Liu, and James M. Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *European Conference on Computer Vision (ECCV)*, September 2018. 2

[29] Yanghao Li, Naiyan Wang, Jianping Shi, Xiaodi Hou, and Jiaying Liu. Adaptive batch normalization for practical domain adaptation. *Pattern Recognition*, 80:109–117, 2018. 2, 5, 6

[30] Ji Lin, Chuang Gan, and Song Han. TSM: Temporal shift module for efficient video understanding. In *International Conference on Computer Vision (ICCV)*, October 2019. 2

[31] Mengyuan Liu, Hong Liu, and Chen Chen. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition*, 68:346–362, 2017. 3

[32] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning (ICML)*, 2015. 1, 2, 6, 8

[33] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International Conference on Machine Learning (ICML)*, 2017. 1, 2

[34] Juan Carlos Niebles, Chih-Wei Chen, and Li Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *European Conference on Computer Vision (ECCV)*, 2010. 3

[35] Andrew Owens and Alexei A. Efros. Audio-visual scene analysis with self-supervised multisensory features. In *European Conference on Computer Vision (ECCV)*, 2018. 2

[36] Boxiao Pan, Zhangjie Cao, Ehsan Adeli, and Juan Carlos Niebles. Adversarial cross-domain action recognition with co-attention. *AAAI Conference on Artificial Intelligence*, 2020. 3

[37] Hamed Pirsiavash and Deva Ramanan. Detecting activities of daily living in first-person camera views. In *Computer Vision and Pattern Recognition (CVPR)*, 2012. 1

[38] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *Computer Vision and Pattern Recognition (CVPR)*, 2012. 2

[39] Fan Qi, Xiaoshan Yang, and Changsheng Xu. A unified framework for multimodal domain adaptation. In *ACM Multimedia Conference on Multimedia Conference*, 2018. 2

[40] Hossein Rahmani and Ajmal Mian. Learning a non-linear knowledge transfer model for cross-view action recognition. In *Computer Vision and Pattern Recognition (CVPR)*, 2015. 3

[41] Kishore K Reddy and Mubarak Shah. Recognizing 50 human action categories of web videos. *Machine Vision and Applications*, 2013. 3

[42] Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele. A database for fine grained activity detection of cooking activities. In *Computer Vision and Pattern Recognition (CVPR)*, 2012. 2

[43] Marcus Rohrbach, Anna Rohrbach, Michaela Regneri, Sikandar Amin, Mykhaylo Andriluka, Manfred Pinkal, and Bernt Schiele. Recognizing fine-grained and composite activities using hand-centric features and script data. *International Journal of Computer Vision*, 2015. 1

[44] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *The European Conference on Computer Vision (ECCV)*, 2010. 5

[45] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 6, 8

[46] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Actor and observer: Joint modeling of first and third-person videos. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 3

[47] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision (ECCV)*, 2016. 1, 2

[48] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems (NIPS)*, 2014. 2

[49] Kihyuk Sohn, Sifei Liu, Guangyu Zhong, Xiang Yu, Ming-Hsuan Yang, and Manmohan Chandraker. Unsupervised domain adaptation for face recognition in unlabeled videos. In *International Conference on Computer Vision (ICCV)*, Oct 2017. 2

[50] Sebastian Stein and Stephen McKenna. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2013. 1, 2

[51] Baochen Sun and Kate Saenko. Deep Coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision (ECCV)*, 2016. 1, 2

[52] Yu Sun, Eric Tzeng, Trevor Darrell, and Alexei A Efros. Unsupervised domain adaptation through self-supervision. *arXiv*, 2019. 2

[53] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *Computer Vision and Pattern Recognition (CVPR)*, 2011. 1

[54] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3D convolutional networks. In *Computer Vision and Pattern Recognition (CVPR)*, 2015. 3

[55] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2

[56] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008. 8

[57] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 5

[58] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal Segment Networks: Towards good practices for deep action recognition. In *European Conference on Computer Vision (ECCV)*, 2016. 2, 5

[59] Donglai Wei, Joseph J Lim, Andrew Zisserman, and William T Freeman. Learning and using the arrow of time. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[60] Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *International Conference on Computer Vision (ICCV)*, Oct 2017. 2

[61] Jiaojiao Zhao and Cees Snoek. Dance with Flow: Two-in-one stream action detection. In *Computer Vision and Pattern*

Recognition (CVPR), 2019. 2

[62] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camera style adaptation for person re-identification. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[63] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *European Conference on Computer Vision (ECCV)*, 2018. 2, 3

[64] Fan Zhu and Ling Shao. Enhancing action recognition by cross-domain dictionary learning. In *British Machine Vision Conference (BMVC)*, 2013. 3

[65] Yang Zou, Zhiding Yu, B.V.K. Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *European Conference on Computer Vision (ECCV)*, September 2018. 2